Yahoo! Movies User Ratings and Descriptive Content Information, v.1.0

Ravitheja Bodempudi
Department of Computer Science
University of Bridgeport, Bridgeport, CT

Abstract

A fundamental aspect of rating based systems is the observation process; the process which users choose the movies they rate. Finding user to user similarity is a fundamental component for collaborative filtering. In user to user similarity ratings assigned by two users to a set of items are pairwise compared and averaged is called correlation. In this project I want to show user to user similarity adaptive i.e., we dynamically change the computation depending on the profiles of the compared users and the target movie whose prediction is sought. I evaluate the proposed theory with k-means clustering by grouping similar users which rated similar movies with same rating. i.e., whoever is having same will come under one group.

Introduction

The goal of clustering is to group similar objects into clusters while separating dissimilar objects. Although clustering is a very popular data mining technique which has been used for over 40 years, its objectives and how to evaluate different clustering results is still subject to a lot of controversy. Because reviewers review a relatively small number of products, it is difficult to find enough reviewers with similar reviews to make accurate predictions for every product each user requests. The problem is exacerbated as products are divided into domains so there are fewer product reviews to train the domain-specific weights with. Resolving the data sparsity problem has been the focus of much recommender system work. Although it is widely accepted that domain specific reviewers result in accurate predictions, it has recently been suggested that a mediated advice giver that combines multiple domains of products and holds only a single set of weights for each, user would help alleviate the data sparsity problem.

The Yahoo! Dataset

This dataset contains a small sample of the Yahoo! Movies community's preferences for various movies, rated on a scale from A+ to F. Users are represented as meaningless anonymous numbers so that no identifying information is revealed. The dataset also contains a large amount of descriptive information about many movies released prior to November 2003, including cast, crew, synopsis, genre, average ratings, awards, etc. The dataset may be used by researchers to validate recommender systems or collaborative filtering algorithms, including hybrid content and collaborative filtering algorithms. The dataset may serve as a testbed for relational learning and data mining algorithms as well as matrix and graph algorithms including PCA and clustering algorithms. The size of this dataset is 23 MB.

Problem Definition

The Yahoo! data set consists of usernames with the movie names they rated; Descriptive reviews for movie. Ratings were given on a scale between 1 to 10. After preprocessing the data ratings were comprised to A+ to F. Problem statement is stated below.

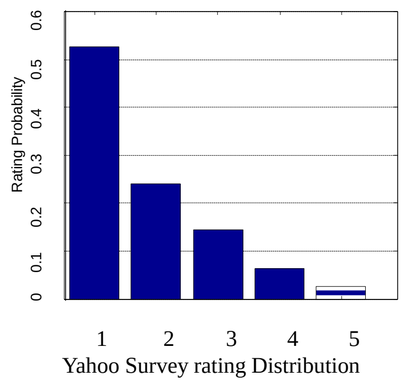
I. Finding the probability of rating a movie for the datasets provided as movie lens, each movie, yahoo! Survey, Yahoo user based.
II. Comparison between Male and Female Movie Rating.

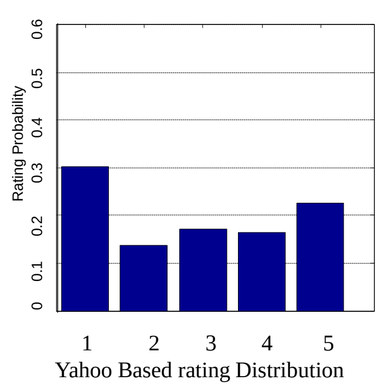
Models and Algorithms

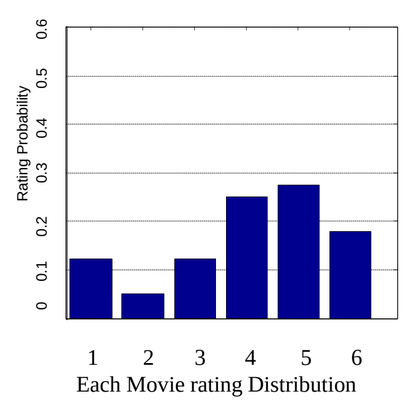
The framework we consider for learning and prediction with non-random missing data follows the basic outline suggested by Little and Rubin. The missing data models we consider capture some properties of a non-random missing data process, but are necessarily simplistic since our aim is to simultaneously estimate the parameters of both the complete data model and the missing data model. We begin by describing three models based on multinomial mixture clustering that incorporate different missing data assumptions. We also briefly describe the baseline matrix factorization and nearest neighbor methods.

Evaluation Analysis

Representing the Movie ratings and the users on either sides of a graph and using k-means clustering defining the groups of people who gave the ratings between A+ to F. By considering a group of movies which were rated or on the basis of no. of movies. Example Evaluation plan is shown below.







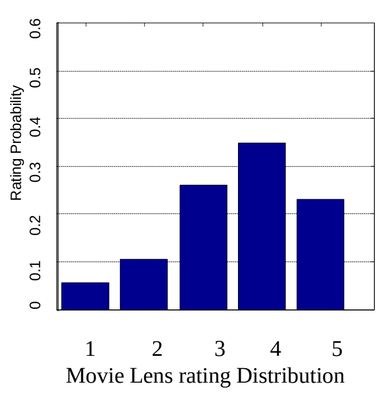


Fig 6.1. Rating Comparison between Male and Female.

Conclusion

A very interesting direction for future research is to consider combining methods that optimize ranking performance, as in the work of, while simultaneously accounting for the presence of non-random missing data. We have argued that the use of randomly selected test items more accurately reflects the tasks of interest: prediction and ranking for items not previously rated by the user. A very interesting direction for future research is to consider combining methods that optimize ranking performance, as in the work of, while simultaneously accounting for the presence of non-random missing data.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/ The MIT Press, 1996.
[2] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.
[3] V. Cherkassky and F. Mulier, "Learning From Data", John Wiley & Sons, 1998.